论文与施引文献引用关联构建及应用*

——解决引文分析中的“卡脖子”问题

王立学

（中国科学院文献情报中心，北京 100190）

摘 要：［目的/意义］研究者需要的论文引用关联需要获知施引文献引用的被引论文，但我们常用的下载数据中却只包含施引文献引用的参考文献。只有建立参考文献条目与被引论文之间的映射，才能让引文分析的研究突破瓶颈、焕发生机。［方法/过程］本文提出组合使用 DOI 匹配和多字段组配的方法，基于对下载论文数据的拆分或解析，本地创建被引论文与施引文献的参考文献条目的关联。［结果/结论］创建论文引用关联是一个基础性的数据处理环节，可以支持多种分析和应用的开展，将为科学计量学开启广阔的发展空间。

关键词：引文分析；施引文献；引用关联；引文数据；文献计量方法

自加菲尔德创立引文索引以来，引文分析在理论、方法和工具等方面取得了全方位发展，在文献关联揭示、论文计量与评价、学科结构分析和科学规律探究等方面已见诸多应用。引文分析的数据基础，是由施引文献引用参考文献而得的论文之间的引用关系，即“被引论文=参考文献→施引文献”。在研究者们的语境里，引用关系一般指的是“被引论文→施引文献”关联，其中默认了已有“被引论文=参考文献”映射关系；不过，商业数据库目前却只能提供“~~被引论文~~=参考文献→施引文献”数据的下载，并不包含前面的映射关系；与此同时，由于缺少便捷可用的工具，普通研究者又很难凭借一己之力将缺失的映射关系补全。最终，论文引用关系数据表现为“线上不好‘整’、线下不完整”，在正常的引用网络中制造出了多处“断点”（见图 1 中的虚线链接），严重阻碍了引文分析在方法研究、指标探索和规律分析等方面向更深入、更科学的方向发展，是引文分析研究与应用中的“卡脖子”问题。

* 本文为国家社科基金项目“中外同学科期刊跨遴选体系联合排序方法研究”（19BTQ190）的研究成果。

作者简介：王立学（ORCID: 0000-0001-9108-7671），副研究馆员，EMAIL: hiwanglixue@163.com。

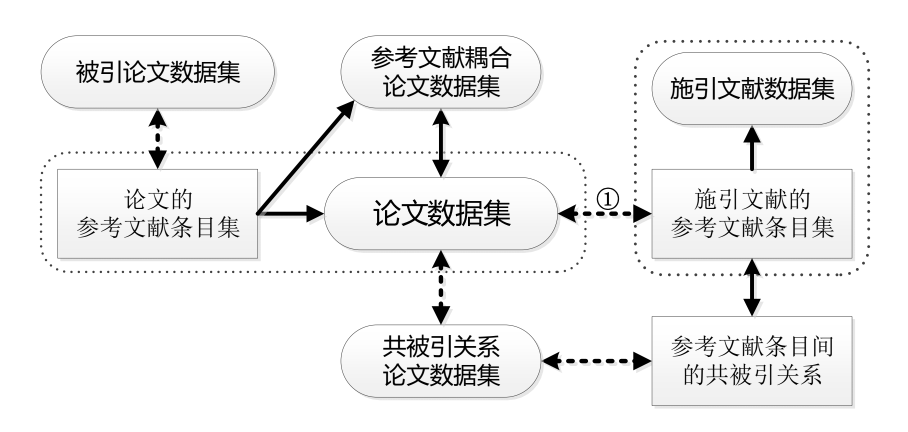


图 1 论文间的引用关系示意图

要突破上述瓶颈、创建“被引论文→施引文献”关联，根本上是建立“被引论文=参考文献”映射关系，即打通图 1 中标记为①的虚线链接。瑞典计量学家 Olle Persson 曾断言将论文和施引文献加以对比会有巨大潜力[1]，我们也可以预见，如果含有完整引用关系的论文数据能够方便地获得，不乏创新性的研究者们必定能在方法研究、指标探索和应用拓展等方面赋予引文分析更大的施展空间，激发出强大的研究活力并开拓出更有价值的应用前景。为此，本研究以 Web of Science（WoS）平台下载的论文数据集及其对应的施引文献数据集（含参考文献）为例，提出本地创建被引论文与施引文献关联的方法，介绍在方法研究和评价指标探索上的现实用例，并进一步探讨所提方案的缺陷和优化措施。

1 现状述评

1.1 论文引用关系数据

多年来，国内外学者使用论文引用关系进行过许多研究，主要有使用引用频次（高被引）进行评估，借助引用关联（共被引、参考文献耦合等）开展引用网络分析，以及将引用结合文献其他特征来尝试关联分析等几种类型。相关研究中，直接使用“参考文献→施引文献”数据的占多数，使用施引文献数据的研究相对较少，构建和使用“被引论文→施引文献”关联的研究则很少。原始引用关系数据方面，多数研究是从 WoS、Scopus、CSCD、CSSCI、CNKI 等数据库中获取，是较为常见的情况；有些研究者有条件使用底层引文数据，如 SCI 光盘数据、WoS BP 数据等，但不具普适性。

使用参考文献数据的研究大致可分为两类：一是单纯使用引用关联数据，即施引文献的参考文献字段，统计分析参考文献信息项中包含的作者、期刊、年份等，或参考文献共被引分析、施引文献的参考文献耦合分析等，但因信息项少而大多仅作为过程结果；二是对具有引用关联的论文进行其他关联分析，即将参考文献与其他内外部特征相结合，开展参考文献与关键词、被引作者与关键词、被引作者与论

文作者、期刊互引、参考文献年份与共被引关系，以及近些年多见的引用内容分析等，所需数据项较丰富，但主要需求仍然可由单一数据集来满足。

使用施引文献数据的研究也主要有两种类型：一类是在数据集层面，单独使用施引文献数据集或组合使用被引论文和施引文献两个数据集进行计量和分析，但并不求被引论文与施引文献的一对一关联；另一类做法是逐篇获得被引论文的施引文献，但因跳过了将被引论文与具体参考文献条目映射的过程，故而多数只能用于小规模数据集，且所得数据集很难有后续通用性。

在关联被引论文与具体参考文献条目的案例中，Olle Persson[2]曾介绍 Bibexcel 软件在此方面的功能，使用按 WoS 参考文献样式组合出的字符串或 DOI 去精确匹配和关联，遗憾的是该软件向用户反馈的只是被引论文与施引文献的对应，而不是具体到对应的参考文献条目；祝清松、冷伏海[3]曾使用 HistCite 软件的“本地被引次数”（LCS）功能将数十篇被引论文关联到施引文献的参考文献条目，该软件适于分析单个数据文件工作，且需较多的后续人工操作才能导出结果和使用；秦晓慧、乐小虬[4]曾提到使用自编 Java 程序构建以单篇论文为核心的论文前后向引用网络，但并未展开描述关联构建细节。

1.2 引用关系数据工具

图情领域中尚未见到专门用于处理引用关系数据的工具，但多数文献计量分析工具或平台都具有一定的引用关系统计功能。论文和报道中多有提及的工具有十多种，包括 Citespace、Bibexcel、VOSviewer、CitNetExplorer、Sci2、SciMAT、HistCite、TDA、SATI、ItgInsight、Bicomb、RefViz、bibliometric.com、COOC 等，探究后发现：现有工具或平台均不能友好地支撑用户自建本地论文引用数据库，实际上它们主要是面向普通用户的数据统计应用，功能重心并不是面向专业用户的基础数据操作；HistCite 和 Bibexcel 能够在一定程度上支持被引论文与施引文献的关联，但操作繁琐且功能有限；Citespace 软件的共被引分析，实为针对数据集内的所有参考文献条目开展。需要说明是，另有多种 R、Python 或其他编程语言环境下的工具包可以进行文献计量分析，限于时间和精力，本文并未进行扩展和探索。

Bibexcel 软件[2]使用“Citations among docs”功能来关联被引论文与施引文献，共有两种方式：一是多字段匹配功能（Make citation links among WoS-records），按照 WoS 参考文献的字段样式，从每条论文数据中抽取所需信息组合成参考文献样式，去数据集中匹配；二是使用 DOI 匹配（Make citation links based on DOI）。运行功能后可获得两列编号：分别是施引文献和被引论文在数据文件中顺序号；使用“Add field to units”功能，可以分别往该列表中添加施引文献的 ID（WoS 数据中的 UT 字段）和被引论文的 ID，由此可得施引文献与被引论文的引用关系。经过测试，Bibexcel 软件的多字段匹配与 DOI 匹配所获引用关系数量不一致，在实际使用中可将二者取并集。软件设计目的并非支撑用户构建本地分析数据库，不太能够胜任创建论文引用关系数据库的需求。

HistCite 软件具有“本地被引次数”（LCS）功能，可以得到施引文献集[3]。该软件同样采用多字段匹配模式，“精确匹配”到施引文献的参考文献条目；不支持使用 DOI 字段的匹配。使用软件中的导出功能，将记录按照“LCS”排序并记录下每篇论文的顺序号，然后点击每条记录后面的 LCS 数值即可打开本地施引文献列表，并可以导出为 CSV 格式。不过，后续还需要根据论文的顺序号补充上每篇被引论文的 ID，才能够获得数据文件中被引论文与施引文献的对应关系。该软件一次只能处理一个 TXT 文档，且需要较多的人工操作，故只适用于小数据量的应用。

从操作上来看，Bibexcel 和 HistCite 基于多字段组合后精确匹配的字符串，对作者姓名的拼写形式极为敏感，稍有不同即无法匹配；Bibexcel 匹配的结果仍然要手动关联被引论文与施引文献的参考文献条目；HistCite 匹配施引文献的结果还需要按照被引论文的顺序号去手动对应，总体而言并不能满足实用需求。

2 论文引用关联构建方法

引文分析研究在理论探索、方法创新和应用拓展等方面均离不开基础引用关系数据的支撑。为了获得所需的论文引用关联，我们针对从 WoS 分别下载的论文数据集及其对应的施引文献数据集，经过研究和探索提出组合使用“DOI 匹配”和“多字段组配”的关联方法（见表 1），将被引论文匹配到施引文献的参考文献条目（被引论文=参考文献），从而实现为本地数据集创建基础引用关系的功能，进而支撑数据库层面的各种分析和统计操作。

表 1 期刊论文引用关联匹配依据

关联匹配依据		使用方式					
DOI 匹配		有则优先					
多字段组配		M1	M2	M3	M4	M5	M6
①	一作姓名		√	√	√		
	一作姓氏	√					
②	出版年	√	√	√	√		
③	期刊名称	√	√	√	√		
④	出版卷	√	无	√	√	√	无
⑤	出版期	√	√	无	√	无	√
⑥	起始页	√	√	√	无	无	无

DOI 匹配并不复杂，但需经过两个预处理步骤才可使用：一是字符规范，可以部分纠正引文数据在 OCR 环节引入的字符识别错误；二是为了能够匹配参考文献中所含 DOI，需要逐条对参考文献条目中的信息进行解析，判断其中 DOI 的完整性和数量。有的数据库中可见到一条参考文献条目标记有多个 DOI 的情况，则不能使用这种情形下的 DOI 信息进行引用关联创建。

多字段组配是将被引论文的相关信息按照数据源的特色样式组配起来，去与解

析后重新组配的施引文献的参考文献匹配。如要考虑多来源数据集合并使用的需求，则还需将所有参考文献条目都进行精确解析，并定义一个统一的多字段组配规则，本文采用的是表 1 中的 M1、M2 和 M3。经过摸索，在多字段组配中有四项内容需作进一步探讨：论文页码、第一作者姓名和姓氏的使用场景及理由、来源期刊名称，以及多来源数据融合场景下的处理。

⑴**论文页码**。页码是有效区分论文的关键信息之一，尤其是在其他组配字段不太完整的时候，页码信息的重要性就更为凸显。若论文缺少页码信息，则通过多字段组配的方式创建的论文引用关联，很可能会错误地匹配上无关论文，此种情况必须施加额外校验；若论文页码存在，只需要使用起始页即可。

⑵**第一作者姓名的处理**。厘清同一作者姓名的不同拼写方式一直是一个难题。从 WoS 数据库的实际数据来看，部分中国学者的拼音姓名存在姓氏和名字颠倒，或是两个字的名字仅仅被标记为一个首字母（如“Qian, Xuesen”被标记为“Qian, X.”）等情况；外国学者则主要存在由是否写全中间名而带来的拼写形式差异。因此，在兼顾容错性和准确度的情况下，我们在其他组配字段完整时仅使用作者的姓氏，这样能尽可能多地创建论文引用关联（如表 1 中的 M1）；在缺少卷或期时使用作者的姓名缩写（如表 1 中的 M2 和 M3），以降低出错的概率。

⑶**来源期刊名称**。不同的数据库对期刊名称的标记略有差别，有的使用简称（如 WoS 数据），有的使用全名（如 Scopus），国内中文期刊名称则基本上都是全名。为了处理多来源数据时匹配，应创建并维护期刊信息表，以便于简称、全名的映射。

⑷**多来源数据融合场景**。多来源数据时的论文引用关联构建，需要识别并标记各个来源中共有的论文记录，然后将各个来源的参考文献样式加以统一，并结合使用论文 DOI。需要指出的是，WoS 数据库的参考文献条目中并不包含“出版期”信息，其规则无法区分只有期数不同的多篇论文（如 DOI 为 10.7500/AEPS20170601011 和 10.7500/AEPS20170120004 的这两篇），故不能用作多来源数据场景下的统一样式。

经过上述处理之后，就可以实施论文记录与参考文献条目的匹配，进而在本地数据库中创建论文引用关联。为了验证实际效果，我们随手选了一个来自 WoS 的含参考文献的数据文件，分别用 HistCite、Bibexcel 和本文所提关联构建方法加以实验，结果为：HistCite 获得 428 对关系，Bibexcel 共获得 655 对关系，本文方法获得了 592 对关系。经人工校验发现：HistCite 虽然匹配关系数量少但无错误；Bibexcel 有 75 对关系匹配错误，但错误原因却难以判断；HistCite 中有 3 对关系 Bibexcel 未能匹配上；本文方法所获结果数量比 HistCite 多了 164 对关系，比 Bibexcel 多了 12 对关系。

3 论文引用关联的应用

创建论文引用关联后，研究者的可用数据将从孤立的论文数据集拓展为关联的“被引论文→施引文献”数据集，进而能够循着论文引用链条开展前后向直接、间接引用分析以及结合文献特征的多角度关联分析。被引论文和施引文献关联后的部

分潜在应用见表 2。受表格维度的限制，表中只能简单展现出两个字段关联的潜在应用，若使用更多字段则有望开拓出更多可能。

表 2 论文与施引文献关联的应用潜力

论文 / 施引文献	作者（第一/通讯）	作者机构	关键词	学科	参考文献	被引次数
作者（第一/通讯）	作者自引/互引/直接引用	作者引用机构	作者引用主题	作者引用学科	作者潜在知识基础	作者引用模式
作者机构	作者的机构影响力	机构自引/互引/直接引用	机构引用主题	机构引用学科	机构潜在知识基础	机构引用模式
关键词	作者的主题影响力	机构的主题影响力	主题关联或演化	学科主题演化	潜在主题关联、知识基础	主题热度和持续度
学科	作者的学科影响力	机构的学科影响力	学科主题演化	学科直引/互引	学科潜在知识基础	学科引用模式
出版年	责任作者被引演化和趋势	主导机构被引演化和趋势	主题被引演化和趋势	学科被引演化和趋势	知识流动速度	被引时间分布及趋势
期刊	作者的期刊影响力	机构的期刊影响力	主题的期刊影响力	学科的期刊影响力	期刊潜在知识基础	加权被引统计
参考文献	作者共被引、耦合	机构共被引、耦合	主题共被引、耦合	学科共被引、耦合	论文被引规律、论文与参考文献共被引	被引来源精确统计
被引次数	论文被引规律	论文被引规律	主题热度和持续度	学科热度	论文被引规律	论文被引规律

基于厘清的论文引用关系，我们已经开展了几种现实应用，例如多种规则的自引统计、学术影响贡献度分析、数据集内共被引分析以及使用加权被引评价论文等。

3.1 自定义规则的自引统计

论文收录引证工作一直是各图书馆的传统业务，经常在项目结题、奖项申报和人才评选等场合提供客观数据的支撑。在统计论文被引数据时，经常需要单列或排除自引次数，但自引的认定可以是作者自引、合著者自引或者机构自引等规则，并且可能与数据库平台的规则不一致，故而实际工作中常常需要人工判断自引次数。

在构建论文论文引用关联之后，清晰掌握论文的每一次被引来源，结合对论文的作者姓名规范、署名情况梳理、隶属机构拆分等数据解析工作，可以区分出来自国内外、机构内外、特定作者群体内外的引用以及互引，也就能方便地统计出各种规则下的自引次数。因论文自引统计的原理较为明确，具体细节在此不作赘述。

3.2 学术影响贡献度分析

高校和科研院所时常需要统计科研论文情况，可能涉及论文的数量统计、本单

位署名情况、被引次数等，更深入的会分析责任人及其二级机构、被引情况和学术影响力分布等内容。以往的被引统计，只能使用数据库平台给出的累积总被引次数和其他统计值，如“破四唯”之前被经常使用的 WoS 被引次数和 ESI 全球前 1%学科指标。不过，这种统计只能限定统计对象的发文年而不能按需设定被引年范围，例如：要了解本单位论文的“近两年情况”，只能看近两年发表论文的情况，却很难分析本单位历年论文在近两年中的被引情况。

基于本地化的论文引用关系，能够根据需要对来自数据库平台的被引数据进行进一步限定和操作。对此，我们主要探索了两种类型的应用：一种是按施引年份来统计论文被引情况，以揭示更具时效性的学术影响力分布，如“某单位论文的近两年被引情况”；二是根据 ESI 的统计规则来限定施引文献的期刊范围和文献类型，统计出特定单位已进入和未进入全球前 1%学科的历次表现及变化趋势，以及内部二级机构在被引方面的贡献度和表现特征，用以辅助科研单位的学科发展规划。

3.3 数据集内共被引分析

前些年，学者们在开展共被引分析时广泛使用 Citespace 软件。不过，该软件的共被引结果取决于使用者所用的基础数据类型：若为目标论文数据集，则所得结果为该论文集引用的参考文献之间的共被引关系；若为施引文献数据集，则结果为包含目标论文和其他被引参考文献的共被引关系。上述两种结果，均不是使用者实际期望的目标论文之间的共被引关系。

因此，通过将被引论文关联到施引文献的参考文献条目，能够轻松获得特定论文集合的共被引关系，支持学者们针对任意的自定义领域开展共被引分析，进而可以利用文献计量手段分析领域的知识基础、领域前沿，并将共被引关联拓展应用到学者、期刊、单位、主题等其他分析对象。

3.4 使用加权被引评价论文

在网络搜索引擎领域，Google 凭借 PageRank 排名算法在创建初期快速赢得了用户认可，其灵感正是来自于文献计量中的论文引用。网页没有可供参考的初始权重，而且存在闭环超链接的情况，故 PageRank 使用随机路径来计算每个网页的权重；与之不同的是，论文引用只能由新到旧单向发展，而且绝大多数的期刊论文都具有天然的初始权重——期刊影响因子。

在当前“破四唯”的评价导向下，社会各界普遍地不再片面强调被引次数的高低、不再粗暴地以刊评文。在这一背景下，我们尊重许多学科都有广受认可的优秀期刊的客观实情，将施引文献的期刊影响因子用作施引权值，来取代以往不考虑施引源学术影响高低的被引计数，从而获得了论文的加权被引次数（详情将另行撰文论述）。由此带来了两个变化：一是加权被引从源头上增加了学术影响力的内涵，一定程度上更能反映论文的实际学术影响；二是有望借助提高有效被引的门槛，来降

低一部分不规范引用行为的发生。论文加权被引的做法，三年来一直为北京市科协举办的“北京地区广受关注学术成果”的遴选及系列报告会提供定量支撑。

4 问题与展望

以可便捷获取的常规论文数据为基础，本文提出组合使用“DOI 匹配”和“多字段组配”的论文引用关联方法，用于精确创建本地数据集内的论文引用关系，将为专业人员开展引文分析方面的方法研究、指标探索和评价应用提供重要的基础条件。以往的分析受制于参考文献条目包含信息太少，如今扩展为带有丰富内外部特征的施引文献与被引论文，为基于引用链条开展的单个或混合式文献特征关联分析打开了探索的大门。

为了更好地使用创建引用关联后的基础数据，我们也必须清晰地了解当前方案中可能存在的问题。首先是要了解数据中特殊情况的存在，如只有卷号或期号（WoS 参考文献中无期号）但无页码的情况，在进行多字段组配时需要在错配和漏配之间作出评估；WoS 下载数据的参考文献条目中存在有标记多个 DOI 的情况，并不能用作绝对稳妥的匹配手段。再者，基础数据来自单个数据源和多个数据源是两种截然不同的情况，单一来源时要按其原本著录的信息进行多字段组配，使用修正后的信息会导致匹配不上；但多个数据源时则需要使用修正后的信息进行多字段组配。更进一步，是逐渐提高所用方案的准确性，大致有三个阶段：一是简单地照原样拆分各字段后使用，二是按规则解析各字段后带一定校验的使用，三是维护和使用多种词表进行信息规范。

通过探索，HistCite 多年未更新且匹配有遗漏、Bibexcel 存在匹配错误的情况且难以控制，最适合的方案是组合使用多字段组配和 DOI 匹配的方法，我们建议相关人员在情况允许的情况下逐步追求匹配准确性的提高。如果图情领域学者能够方便地构建带有准确引用关系的基础论文数据开展研究和探索，必将为科学计量学开启更为广阔的研究和应用领域。

参考文献

[1] PERSSON O. Exploring the analytical potential of comparing citing and cited source items[J]. Scientometrics, 2006,68(3): 561-572.
[2] PERSSON O,DANELL R,SCHNEIDER JW. How to use Bibexcel for various types of bibliometric analysis[M]. // STRÖM F. Celebrating scholarly communication studies: A festschrift for Olle Persson at his 60th birthday. Leuven: International society for scientometrics and informetrics, 2009: 9-24.
[3] 祝清松,冷伏海. 基于引文内容分析的高被引论文主题识别研究[J]. 中国图书馆学报, 2014,40(209): 39-49.
[4] 秦晓慧,乐小虬. 面向单篇文献引文网络的主题来源与走向追踪[J]. 现代图书情报技术, 2015 (9): 52-59.

Mapping between Paper and Citing Papers to Enable Citation Analysis: A Solution to A Bottleneck Issue in Citation Dataset

Li-Xue WANG

(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [Purpose /significance] Citation Analysis has been suffered all the time from crippled citation data without Links between cited Paper and Citing Papers in hand. Database vendors allow only Reference Items download while keep the real needed Links in the backstage. Scholars have to find ways to establish the Links, so they can break through the bottleneck. [Method /process] We propose a method of combining DOI matching and multi-field matching to create local mappings between cited Paper and Reference Items of Citing Papers, based on the splitting or parsing of downloaded citation data. [Result /conclusion] As a fundamental step in data processing, mapping Links will support a variety of analysis and applications, and will open up a broad space for scientometrics.

Keywords: Citation Analysis; Citing Paper; Citation Link; Citation Data; Bibliometrics Method